Incremental Ontology-Based Extraction and Alignment in Semi-Structured Documents

Mouhamadou Thiam, Nacéra Bennacer Seghouani, Nathalie Pernelle, Moussa Lô

Incremental Ontology-Based Extraction and Alignment in Semi-Structured Documents

Mouhamadou Thiam [1,3], Nacéra Bennacer [2], Nathalie Pernelle [1], and Moussa Lô [3]

[1]LRI, Université Paris-Sud 11, INRIA Saclay Ile de France
2-4 rue Jacques Monod, F-91893 Orsay Cedex, France
[2]SUPELEC, 3 rue joliot- curie, F-91192 Gif-sur-Yvette cedex, France
[3]LANI, Université Gaston Berger, UFR S.A.T, BP 234 Saint-Louis, Sénégal
{mouhamadou.thiam,nathalie.pernelle}@lri.fr
{nacera.bennacer}@supelec.fr
{lom}@ugb.sn
http://www.lri.fr/~thiam

Abstract. *SHIRI* [1] is an ontology-based system for integration of semi-structured documents related to a specific domain. The system's purpose is to allow users to access to relevant parts of documents as answers to their queries. *SHIRI* uses RDF/OWL for representation of resources and SPARQL for their querying. It relies on an automatic, unsupervised and ontology-driven approach for extraction, alignment and semantic annotation of tagged elements of documents. In this paper, we focus on the *Extract-Align* algorithm which exploits a set of named entity and term patterns to extract term candidates to be aligned with the ontology. It proceeds in an incremental manner in order to populate the ontology with terms describing instances of the domain and to reduce the access to extern resources such as Web. We experiment it on a HTML corpus related to call for papers in computer science and the results that we obtain are very promising. These results show how the incremental behaviour of *Extract-Align* algorithm enriches the ontology and the number of terms (or named entities) aligned directly with the ontology increases.

Key words: Information Extraction, Semantic Annotation, Alignment, Ontology, Semi-structured documents, OWL, RDF/RDFS

1 Introduction

Information available on the Web is mostly in HTML form and thus is more or less syntactically structured. The need to automate these information processing, their exploitation by applications and their sharing justify the interest that research carries on the semantic Web. Because of the lack of semantic, the querying over these resources are generally based on keywords. This is not satisfying because it does not ensure answer relevance and the answer is a whole

[1] Système Hybride d'Intégration et de Recherche d'Information, Digiteo labs project

document. The annotation of web resources with semantic metadata should allow better interpretation of their content. The metadata semantics are defined in a domain ontology through domain concepts and their relations. Nevertheless, manual annotation is time-consuming and the automation of annotation techniques is a key factor for the future web and its scaling-up.

Many works belonging to complementary research fields such as machine learning, knowledge engineering and linguistics investigate the issue of annotation of such documents. Some works are based on supervised approaches or on the existence of structure models in the input documents as in [7], [8], [10] or in text as in [3], [12]. Generally, the assumed hypotheses are incompatible with the heterogeneity and the great number of documents. Now, one information may appear in different kinds of structure depending on the document forms. Some unsupervised approaches are specialized in structured parts such as tables [15]. Moreover, one document may contain both structured and unstructured parts.

Except for named entities, instances are often drowned in text, so they are not easily dissociable. Even advanced Natural Language Processing techniques often adapted to very specific corpora could not succeed.

Named Entities Recognition (NER) aims to locate and classify elements in text into predefined categories such as the names of persons, organizations, locations, dates, etc. Some unsupervised Named-entity recognition systems are based on lexical resources ([9]), or on lexical resources built thanks to data available on the web ([12], [2]). Some approaches use the Web as a possible corpora to apply pattern and find terms to annotate a named entity of a resource [3]. Because this method is time-consuming, it has to be applied when other strategies fail.

The automation of heterogeneous documents annotation can also be based on terms that describe concepts that are not named entities. The different extraction techniques can be categorized as linguistic, statistic or hybrid ([14], [13]).

Once a term or a named entity is extracted, it has to be compared to the set of terms that belongs to the Ontology (concept labels or named entities). Similarity measures that can be used to estimate a semantic similarity between named entities or terms have been extensively studied ([6]).

SHIRI[1] can be introduced as an ontology-based integration system for semi-structured documents related to a specific domain. The system purpose is to allow users to access to relevant parts of HTML documents as answers to their queries. *SHIRI* uses RDF/OWL standard W3C languages for representation of resources and SPARQL for their querying. The system relies on an automatic, unsupervised and ontology-driven approach for extraction, alignment and semantic annotation of documents tagged elements. The extraction of term candidates to be aligned with the ontology relies on a set of named entity and term patterns. It proceeds in an incremental manner in order to populate the ontology with terms describing domain instances and to reduce the access to extern resources such as Web. The annotation of these terms is associated to tagged element of the HTML document (named structural unit) [1]. Actually, terms are generally not located in an accurate manner and may be drowned in a same structural unit. In this paper we focus on the algorithm defined for the

extraction and the alignment named *Extract-Align* algorithm. We experiment and validate it on a HTML corpus related to call for papers in computer science and the results that we obtain are very promising. These results show how the incremental behaviour of *Extract-Align* algorithm enriches the ontology and how the number of terms (or named entities) aligned directly with the ontology increases. In section 2, we detail the extraction and alignment approach. In section 3, we present the results of the experiments made on a corpus related to call for papers. In section 4, we conclude and give some perspectives.

2 Incremental and Semantic Alignment Approach

In this section, we focus on the terms extraction and their alignment with the ontology. The extraction method applies a set of patterns to extract term candidates. It distinguishes the named entity patterns and the term patterns. The term candidates are to be aligned with the concepts of the domain ontology. This alignment is either directly done with the ontology or indirectly thanks to the Web. The ontology is then populated with the aligned terms that are exploited for the next alignments.

2.1 Ontology description

Let $\mathcal{O}(\mathcal{C}, \mathcal{R}, \preceq, \mathcal{S}, \mathcal{A}, \mathcal{L}_{\mathcal{EX}})$ be the domain ontology where $\mathcal{C}$ is the set of concepts, $\mathcal{R}$ is the set of relations between concepts, $\preceq$ denotes the subsumption relation between concepts and between relations. $\mathcal{S}$ defines the domain and the range for each relation and $\mathcal{A}$ is a set of axioms and rules defined over concepts and relations.

$\mathcal{L}_{\mathcal{EX}}(\mathcal{L}, \mathcal{T}, prefLabel, altLabel, hasTerm, hastermNe)$ defines the set $\mathcal{L}$ of concept labels and the set $\mathcal{T}$ of terms or named entites describing the concepts of the domain. Each concept $c \in \mathcal{C}$ is related to a preferred label via $prefLabel$ property and to alternate labels via $altLabel$ [2] belonging to $\mathcal{L}$. Each concept $c \in \mathcal{C}$ is related to terms via $hasTerm$ property and to named entities via $hastermNe$ belonging to $\mathcal{T}$. We assume that the sets $\mathcal{L}$ and $\mathcal{T}$ are initialized respectively by a set of labels and a set of terms selected by the domain expert.

Example 1. Labels and terms selected for the $Topic$ concept c of computer science domain include the following:
prefLabel(c, 'Topic'), altLabel(c, 'field'), altLabel(c, 'area'), altLabel(c, 'theme'), hasTerm(c, 'communications protocol'), hasTerm(c, 'data encryption'), hasTerm(c, 'information'), hasTerm(c, 'object-oriented programming language')
The set of terms $\mathcal{T}$ is enriched by extracted terms as documents are processed. Since this enrichment is automatic, some terms may be irrelevant, that's why we distinguish them from labels. If the expert decides to validate the ontology, it is possible that some of them become labels.

[2] Properties defined in SKOS: Simple Knowledge Organization System

2.2 Extract-Align Algorithm

The *SHIRI* extraction and alignment approach proceeds in an incremental manner. Each Extract-Align invocation processes a subset of documents. More precisely, at each invocation, the algorithm is applied to a subset of documents D belonging to the same domain, to the ontology of this domain $\mathcal{O}$, to a set of patterns P, to a set $Processed$ of terms handled in previous steps. The algorithm distinguishes two types of patterns : syntactic named entity patterns and syntactic term patterns. These two types of patterns are used to extract a set of term candidates denoted $\mathcal{I}$ (see example in table 1). Each $t \in \mathcal{I}$ is identified by the sequence of the numbered words according to their occurrence order in the document. These terms are to be aligned wih the set of labels $\mathcal{L}$ and the set of terms $\mathcal{T}$ defined in the ontology $\mathcal{O}$.

At each step, the algorithm attempts to directly align terms of $\mathcal{I}$ with the ontology, otherwise by using the web. Besides, each step enriches the set $\mathcal{T}$ of domain terms and named entities, so the number of web invocations should be reduced when the next documents will be processed. That is also the reason why the set of unaligned processed terms are kept in $Processed$.

The function $alignTerm(t)$ is applied to each $t \in \mathcal{I}$ and returns a set of concepts $C_t \subset \mathcal{C}$ if it succeeds. Then, t is added to $\mathcal{T}$ and related to each $c \in C_t$ via $hasTerm$ or via $hastermNe$ relations depending on the matched pattern (see example below). The invoked $alignTerm(t)$ function uses similarity measures that are appropriate to compare two named entities or two terms.

The unaligned terms are submitted to the Web like in CPankow approach [3]: lexico-syntactic Hearst patterns for hyponymy [5] are used to construct queries containing the unaligned term t. These queries are submitted to a search engine in order to find a set of label candidates L_t. For each $l \in L_t$, the function $webAlign(l)$ is applied and returns a set of concepts $C_t \subset \mathcal{C}$. If $webAlign(l)$ succeeds, then, l and t are added to $\mathcal{T}$. t is related to each $c \in C_t$ via $hasTerm$ or via $hastermNe$ relations depending on the matched pattern. l is related to each $c \in C_t$ via $hasTerm$ relation. Since l is extracted automatically it is considered as a term.

In addition, the term candidates $\mathcal{I}$ are also processed in an incremental manner from the longest to the shortest. We assume that a term is more precise and meaningful than the terms it contains. For example *distributed databases* is more precise than *databases*. But for a term such as *Interoperability of data on the Semantic Web*, the alignment will fail very probably. We denote a term of length k occurring at position i in the document as a sequence of k words: $t_i^k = w_i w_{i+1} ... w_{i+k-1}$, where w_{i+j} denote the word at position $i+j$, $j \in [0, k-1]$. We note $\mathcal{I}^k = \{t_i^k, i \in [1, N]\}$ the set of extracted terms of length k varying from len to 1 (len is the maximal length).

At iteration k, the algorithm proceeds terms of $\mathcal{I}^k$ and $\mathcal{I} = \bigcup_{i=1}^{k} \mathcal{I}^k$. We say that $t_{i_2}^{k_2}$ is included in $t_{i_1}^{k_1}$ if $k_2 < k_1$ and $i_2 \in [i_1, i_1 + k_1 - 1]$. When the system aligns a term $x \in \mathcal{I}^k$ then $\forall y \in \bigcup_{i=1}^{k-1} \mathcal{I}^i$ such that y is included in x, y is deleted from $\mathcal{I}$.

Example: Given the text in table 1 and the two patterns $P_t^1 = JN$ and $P_t^2 = N$ where J denotes an adjective and N a name, the extracted terms are the following: In this example, $\mathcal{I}^1 = \{\text{Areas}_{71}, \text{databases}_{76}, \text{intelligence}_{79}, \text{workshop}_{81},$

Original Text	**Extracted Terms**
..Areas$_{71}$ of$_{72}$ interest$_{73}$ are$_{74}$ distributed$_{75}$ databases$_{76}$ and$_{77}$ artificial$_{78}$ intelligence$_{79}$. The$_{80}$ workshop$_{81}$ SEMMA$_{82}$ focuses$_{83}$ also$_{84}$ on$_{85}$ databases$_{86}$. Intelligence$_{87}$ areas$_{88}$..	..[Areas$_{71}$] of$_{72}$ interest$_{73}$ are$_{74}$ [distributed$_{75}$ [databases$_{76}$]] and$_{77}$ [artificial$_{78}$ [intelligence$_{79}$]]. The$_{80}$ [workshop$_{81}$] SEMMA$_{82}$ focuses$_{83}$ also$_{84}$ on$_{85}$ [databases$_{86}$]. [Intelligence$_{87}$] [areas$_{88}$]..

Table 1. An example of extracted terms

databases$_{86}$, Intelligence$_{87}$, areas$_{88}\}$ and $\mathcal{I}^2 = \{$distributed$_{75}$ databases$_{76}$, artificial$_{78}$ intelligence$_{79}\}$. The terms [*distributed*$_{75}$ *databases*$_{76}$] and [*artificial*$_{78}$ *intelligence*$_{79}$] are aligned with the concept [*Topic*]. So, we delete *databases*$_{76}$, *Intelligence*$_{79}$ from $\mathcal{I}^1$.

Three kinds of outputs result from the Extract-Align invocation: (1) rdf triples which enrich the ontology with terms or named entities describing the concepts ($hasTerm$ and $hastermNe$ relations), (2) rdf triples refering the structural units of the documents, the concepts these units contain ($containInstanceOf$) and the values of corresponding terms or named entities ($hasValueInstance$) and (3) the set of all processed terms.

3 Validation of Extract-Align Algorithm

Let $\mathcal{O}$ be the domain ontology of *Call for Papers for Computer Science Conferences*. The named entities are the events (i.e. conferences, workshops), the persons, their affiliations (i.e. team, laboratory and/or university), and the locations (university, city or country) of the events. Each concept is described by a preferred label and a set of alternative labels. For example, *scientist* and*people* are related to the *Person* concept. The expert has also exploited Wordnet to select a set of 353 domain terms such as {*Communications protocol, data encryption, information, ...*} that are related to *Topic* concept via $hasTerm$ property. The corpus we collect is composed of a set of 691 HTML documents (250542 words after pre-processing).

Named entities are automatically extracted from the document collection using Senellart specialized technique [2] which exploits DBLP (Digital Bibliography and Library Project) to identify accurately person names and dates. We also use the set of C-Pankow syntactic patterns to extract other named entities instances. Terms are extracted using the patterns defined in [4].

To retrieve web label candidates, a set of queries is constructed for each term or named entity. The queries are constructed like in C-Pankow approach

using : hearst patterns, copula and definites (noun phrase introduced by the definite determiner *The*). The web labels are selected when the confidence measure value is over 0.2. For the alignment of the term candidates or web labels, we use Taxomap tool [11]. Its aim is to discover correspondences between concepts of two taxonomies. It relies on terminological and structural techniques applied sequentially and performs an oriented alignment from a source ontology to a target ontology. We only exploit the terminological strategies of Taxomap i.e. syntactic-based similarity measures applied on concept labels and terms (term inclusion, n-gram similarity,...). For the alignment of named entities, the similarity is strictly the equality between terms.

For example, in our experiments, the term *reinforcement learning* is directly aligned with the $Topic$ concept thanks to the term *learning*. The term *World Wide Web* has not been aligned directly with $\mathcal{O}$. One of the web label candidate is *information ressources*, Taxomap aligns it with *information* term related to $Topic$ concept. *Reinforcement learning, information ressources, World Wide Web* are then added to $\mathcal{T}$ and related to $Topic$ concept via $hasTerm$ property.

	Named Entity Patterns					Term Patterns		
	Aligned With $\mathcal{O}$	Using the Web	Precision	Precision with incomplete NE	Recall	Aligned With $\mathcal{O}$	Precision	Recall
Affiliation	0	1317	84.18%	86.07%	70.83%	165	96.97%	91.95%
Location	0	1097	98.53%	99.02%	91.86%	143	80.42%	78.77%
Person	745	362	89.47%	90.85%	79.5%	206	63.59%	59.01%
Event	0	741	64.35%	86.13%	84.47%	80	65.00%	65.00%
Date	456	0	97.58%	97.58%	74.17%	-	-	-
Topic	-	-	-	-	-	276	65.58%	59.34%

Table 2. Results for Named Entity and Term Pattern

Table 2 shows the results we obtain for named entity patterns. We present the precision and recall measures for named entites that are aligned either directly or using the Web label candidates. Since the granularity of the *Shiri-Annot* annotation is the structural unit, we consider that a named entity is incomplete but correct if the complete name appears in the structural unit where it is extracted. For instance, *International Conference* is incomplete but the structural unit contains the whole name which is *International Conference on Web Services*. The Web allows to align approximatively 74 % of all the aligned named entities. All the affiliations, events and locations are found thanks to the Web. Furthermore, the table shows that by taking into account incomplete named entities the precision increases especially for events. These named entities are often partially extracted due to their length and their complexity. Table 2 shows that thanks to term patterns, the concept $Topic$ is enriched of 78% de terms. Other named entities have been also found with good precision and recall for affiliations which are often described using complex terms. Table 3 shows that a lot of terms occurs many times since all documents talk about the same domain. Obviously, most

Length	1	2	3	4	5	7	**Sum**
Extracted Terms	101430	32712	17912	5704	966	104	158828
Extracted Terms (distinct)	14413	15806	10020	3797	602	48	44680

Table 3. Number of Extracted Terms by Length

of them are not aligned with the ontology. Moreover, those which are included in aligned terms are not processed.

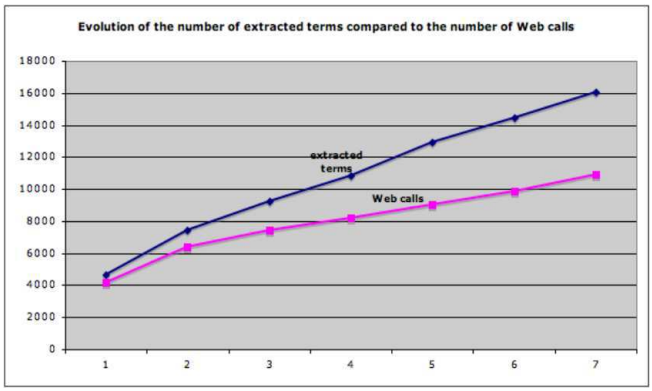

Fig. 1. Extracted terms, Web calls versus the number of documents (by ten)

Figure 1 shows : (1) the evolution of the number of terms extracted according to the number of documents (by ten) (2) the evolution of the number of web calls according to the number of documents (by ten). The results show that the number of Web invocations decreases with the number of processed documents. This explains by the incremental behaviour of Extract-Align algorithm : (1) the more the ontology is populated by new terms, the more a term candidate can be directly aligned and (2) all term web alignements which fail are stored (*Processed* data).

4 Conclusions and Future Works

In this paper, we have presented an automatic, unsupervised and ontology-driven approach for extraction, alignment and semantic annotation of tagged elements of documents. The *Extract-Align* algorithm proceeds in an incremental manner in order to populate the ontology with terms describing instances of the domain and to reduce the access to extern resources such as Web.

We experiment and validate our approach on a HTML corpus related to call for papers in computer science and the results are promising. These results show

how the ontology is enriched and how the number of terms (or named entities) aligned directly with the ontology increases. The constructed ontology can be validated by a domain expert in order to select among the terms those to be removed or those to become concept labels.

A short-term perspective is the exploitation of the annotation model to reformulate domain queries in order to adapt them to the various levels of precision of the annotations. A further perspective is to study how a quality measure can be associated to each annotation triple. We also plan to apply our approach to other domains like e-commerce web sites.

References

1. Thiam M., Pernelle N. , Bennacer N.: Contextual and Metadata-based Approach for the Semantic Annotation of Heterogeneous Documents. ESWC-SeMMA workshop, Tenerife, Spain, 2008.
2. Senellart P. : Understanding the Hidden Web. PHD Thesis, University of Paris 11, December 2007.
3. Cimiano P., Handschuh S., Staab S.: Gimme'The Context : Context Driven Automatic Semantic Annotation With C-PANKOW. WWW conference, 2005.
4. Antti Arppe: Term Extraction from unrestricted Text , the Nordic Conference on Computational Linguistics (NoDaLiDa), 1995.
5. Hearst, M. Marti, A.:Automatic acquisition of hyponyms from large text corpora, Proceedings of the 14th International conference on Computational linguistics, 1992, pages 539–545, France.
6. William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg.: A comparison of string distance metrics for name-matching tasks. In IIWeb, pages 73-78, 2003.
7. Crescenzi V., Mecca G., Merialdo P.: RoadRunner : Towards Automatic Data Extraction from Large Web Sites. Very Large Data Bases Conference (VLDB), 2001.
8. Davulcu H., Vadrevu S. and Nagarajan S.: OntoMiner : Automated Metadata and instance Mining from News Websites. The International Journal of Web and Grid Services (IJWGS), Vol. 1, No. 2, pp. 196-221, Inderscience Publishers, 2005.
9. Borislav P., Atanas K., Angel K., Dimitar M., Damyan O., Miroslav G.: KIM - Semantic Annotation Platform. Journal of Natural Language Engineering vol 10 issue 3-4, Cambridge University Press, pages 375-392, 2004.
10. Baumgartner R. and Flesca S. and Gottlob G: Visual Web Information Extraction with Lixto. The VLDB Journal, pages 119-128, 2001.
11. Hamdi F., Zargayouna H., Safar B., Reynaud C.: TaxoMap in the OAEI 2008 alignment contest, Ontology Alignment Evaluation Initiative (OAEI) 2008 Campaign - Int. Workshop on Ontology Matching, 2008.
12. Etzioni O., Cafarella M., Downey D., Kok S., Popescu A., Shaked T., Soderland S., Weld D., and Yates A. Unsupervised named-entity extraction from the web: An experimental study. Articial Intelligence, 165(1):91134, 2005.
13. R. Navigli, P. Velardi. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites, Computational Linguistics, 30(2), MIT Press, 2004, pp. 151-179.
14. Drouin P. "Term extraction using non-technical corpora as a point of leverage", In Terminology, vol. 9, no 1, p. 99-117,2003.
15. Cafarella M.J., Halevy A., Zhe Wang D. Uncovering the relational web, proceedings of WebDB, Canada, 2008.